The QT21/HimL Combined Machine Translation System

Jan-Thorsten Peter[1], **Tamer Alkhouli**[1], **Hermann Ney**[1], **Matthias Huck**[2], **Fabienne Braune**[2], **Alexander Fraser**[2], **Aleš Tamchyna**[2,3], **Ondřej Bojar**[3], **Barry Haddow**[4], **Rico Sennrich**[4], **Frédéric Blain**[5], **Lucia Specia**[5], **Jan Niehues**[6], **Alex Waibel**[6], **Alexandre Allauzen**[7], **Lauriane Aufrant**[7,8], **Franck Burlot**[7], **Elena Knyazeva**[7], **Thomas Lavergne**[7], **François Yvon**[7], **Stella Frank**[9], **Mārcis Pinnis**[10]

[1]RWTH Aachen University, Aachen, Germany
[2]LMU Munich, Munich, Germany
[3]Charles University in Prague, Prague, Czech Republic
[4]University of Edinburgh, Edinburgh, UK
[5]University of Sheffield, Sheffield, UK
[6]Karlsruhe Institute of Technology, Karlsruhe, Germany
[7]LIMSI, CNRS, Université Paris Saclay, Orsay, France
[8]DGA, Paris, France
[9]ILLC, University of Amsterdam, Amsterdam, The Netherlands
[10]Tilde, Riga, Latvia
[1]{peter,alkhouli,ney}@cs.rwth-aachen.de
[2]{mhuck,braune,fraser}@cis.lmu.de
[3]{tamchyna,bojar}@ufal.mff.cuni.cz
[4]bhaddow@inf.ed.ac.uk rico.sennrich@ed.ac.uk
[5]{f.blain,l.specia}@sheffield.ac.uk
[6]{jan.niehues,alex.waibel}@kit.edu
[7]{allauzen,aufrant,burlot,knyazeva,lavergne,yvon}@limsi.fr
[9]s.c.frank@uva.nl
[10]marcis.pinnis@tilde.lv

Abstract

This paper describes the joint submission of the QT21 and HimL projects for the English→Romanian translation task of the *ACL 2016 First Conference on Machine Translation* (WMT 2016). The submission is a system combination which combines twelve different statistical machine translation systems provided by the different groups (RWTH Aachen University, LMU Munich, Charles University in Prague, University of Edinburgh, University of Sheffield, Karlsruhe Institute of Technology, LIMSI, University of Amsterdam, Tilde). The systems are combined using RWTH's system combination approach. The final submission shows an improvement of 1.0 BLEU compared to the best single system on newstest2016.

1 Introduction

Quality Translation 21 (QT21) is a European machine translation research project with the aim of substantially improving statistical and machine learning based translation models for challenging languages and low-resource scenarios.

Health in my Language (HimL) aims to make public health information available in a wider variety of languages, using fully automatic machine translation that combines the statistical paradigm with deep linguistic techniques.

In order to achieve high-quality machine translation from English into Romanian, members of the QT21 and HimL projects have jointly built a combined statistical machine translation system. We participated with the QT21/HimL combined machine translation system in the WMT 2016 shared task for machine translation of news.[1] Core components of the QT21/HimL combined system are twelve individual English→Romanian translation engines which have been set up by different QT21 or HimL project partners. The outputs of all these individual engines are combined using the system combination approach as imple-

[1] http://www.statmt.org/wmt16/translation-task.html

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 344–355,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

mented in Jane, RWTH's open source statistical machine translation toolkit (Freitag et al., 2014a). The Jane system combination is a mature implementation which previously has been successfully employed in other collaborative projects and for different language pairs (Freitag et al., 2013; Freitag et al., 2014b; Freitag et al., 2014c).

In the remainder of the paper, we present the technical details of the QT21/HimL combined machine translation system and the experimental results obtained with it. The paper is structured as follows: We describe the common preprocessing used for most of the individual engines in Section 2. Section 3 covers the characteristics of the different individual engines, followed by a brief overview of our system combination approach (Section 4). We then summarize our empirical results in Section 5, showing that we achieve better translation quality than with any individual engine. Finally, in Section 6, we provide a statistical analysis of certain linguistic phenomena, specifically the prediction precision on morphological attributes. We conclude the paper with Section 7.

2 Preprocessing

The data provided for the task was preprocessed once, by LIMSI, and shared with all the participants, in order to ensure consistency between systems. On the English side, preprocessing consists of tokenizing and truecasing using the Moses toolkit (Koehn et al., 2007).

On the Romanian side, the data is tokenized using LIMSI's tokro (Allauzen et al., 2016), a rule-based tokenizer that mainly normalizes diacritics and splits punctuation and clitics. This data is truecased in the same way as the English side. In addition, the Romanian sentences are also tagged, lemmatized, and chunked using the TTL tagger (Tufiş et al., 2008).

3 Translation Systems

Each group contributed one or more systems. In this section the systems are presented in alphabetic order.

3.1 KIT

The KIT system consists of a phrase-based machine translation system using additional models in rescoring. The phrase-based system is trained on all available parallel training data. The phrase table is adapted to the SETimes2 corpus (Niehues and Waibel, 2012). The system uses a pre-reordering technique (Rottmann and Vogel, 2007) in combination with lexical reordering. It uses two word-based n-gram language models and three additional non-word language models. Two of them are automatic word class-based (Och, 1999) language models, using 100 and 1,000 word classes. In addition, we use a POS-based language model. During decoding, we use a discriminative word lexicon (Niehues and Waibel, 2013) as well.

We rescore the system output using a 300-best list. The weights are optimized on the concatenation of the development data and the SETimes2 dev set using the ListNet algorithm (Niehues et al., 2015). In rescoring, we add the source discriminative word lexica (Herrmann et al., 2015) as well as neural network language and translation models. These models use a factored word representation of the source and the target. On the source side we use the word surface form and two automatic word classes using 100 and 1,000 classes. On the Romanian side, we add the POS information as an additional word factor.

3.2 LIMSI

The LIMSI system uses NCODE (Crego et al., 2011), which implements the bilingual n-gram approach to SMT (Casacuberta and Vidal, 2004; Crego and Mariño, 2006; Mariño et al., 2006) that is closely related to the standard phrase-based approach (Zens et al., 2002). In this framework, translation is divided into two steps. To translate a source sentence into a target sentence, the source sentence is first reordered according to a set of rewriting rules so as to reproduce the target word order. This generates a word lattice containing the most promising source permutations, which is then translated. Since the translation step is monotonic, this approach is able to rely on the n-gram assumption to decompose the joint probability of a sentence pair into a sequence of bilingual units called tuples.

We train three Romanian 4-gram language models, pruning all singletons with KenLM (Heafield, 2011). We use the in-domain monolingual corpus, the Romanian side of the parallel corpora and a subset of the (out-of-domain) Common Crawl corpus as training data. We select in-domain sentences from the latter using the Moore-Lewis (Moore and Lewis, 2010) filtering method, more specifically its implementation in XenC (Rousseau, 2013). As a result, one third of the initial corpus is removed. Finally, we make a linear interpolation of these models, using the SRILM toolkit (Stolcke, 2002).

3.3 LMU-CUNI

The LMU-CUNI contribution is a constrained Moses phrase-based system. It uses a simple factored setting: our phrase table produces not only the target surface form but also its lemma and morphological tag. On the input, we include lemmas, POS tags and information from dependency parses (lemma of the parent node and syntactic relation), all encoded as additional factors.

The main difference from a standard phrase-based setup is the addition of a feature-rich discriminative translation model which is conditioned on both source- and target-side context (Tamchyna et al., 2016). The motivation for using this model is to better condition lexical choices by using the source context and to improve morphological and topical coherence by modeling the (limited left-hand side) target context.

We also take advantage of the target factors by using a 7-gram language model trained on sequences of Romanian morphological tags. Finally, our system also uses a standard lexicalized reordering model.

3.4 LMU

The LMU system integrates a discriminative rule selection model into a hierarchical SMT system, as described in (Tamchyna et al., 2014). The rule selection model is implemented using the high-speed classifier Vowpal Wabbit[2] which is fully integrated in Moses' hierarchical decoder. During decoding, the rule selection model is called at each rule application with syntactic context information as feature templates. The features are the same as used by Braune et al. (2015) in their string-to-tree system, including both lexical and soft source syntax features. The translation model features comprise the standard hierarchical features (Chiang, 2005) with an additional feature for the rule selection model (Braune et al., 2016).

Before training, we reduce the number of translation rules using significance testing (Johnson et al., 2007). To extract the features of the rule selection model, we parse the English part of our training data using the Berkeley parser (Petrov et al., 2006). For model prediction during tuning and decoding, we use parsed versions of the development and test sets. We train the rule selection model using VW and tune the weights of the translation model using batch MIRA (Cherry and Foster, 2012). The 5-gram language model is trained using KenLM (Heafield et al., 2013) on the Romanian part of the Common Crawl corpus concatenated with the Romanian part of the training data.

3.5 RWTH Aachen University: Hierarchical Phrase-based System

The RWTH hierarchical setup uses the open source translation toolkit Jane 2.3 (Vilar et al., 2010). Hierarchical phrase-based translation (HPBT) (Chiang, 2007) induces a weighted synchronous context-free grammar from parallel text. In addition to the contiguous lexical phrases, as used in phrase-based translation (PBT), hierarchical phrases with up to two gaps are also extracted. Our baseline model contains models with phrase translation probabilities and lexical smoothing probabilities in both translation directions, word and phrase penalty, and enhanced low frequency features (Chen et al., 2011). It also contains binary features to distinguish between hierarchical and non-hierarchical phrases, the glue rule, and rules with non-terminals at the boundaries. We use the cube pruning algorithm (Huang and Chiang, 2007) for decoding.

The system uses three backoff language models (LM) that are estimated with the KenLM toolkit (Heafield et al., 2013) and are integrated into the decoder as separate models in the log-linear combination: a full 4-gram LM (trained on all data), a limited 5-gram LM (trained only on in-domain data), and a 7-gram word class language model (wcLM) (Wuebker et al., 2013) trained on all data and with a output vocabulary of 143K words.

The system produces 1000-best lists which are reranked using a LSTM-based (Hochreiter and Schmidhuber, 1997; Gers et al., 2000; Gers et al., 2003) language model (Sundermeyer et al., 2012) and a LSTM-based bidirectional joined model (BJM) (Sundermeyer et al., 2014a). The models have a class-factored output layer (Goodman, 2001; Morin and Bengio, 2005) to speed up training and evaluation. The language model uses 3 stacked LSTM layers, with 350 nodes each. The BJM has a projection layer, and computes a for-

[2] http://hunch.net/~vw/ (VW). Implemented by John Langford and many others.

ward recurrent state encoding the source and target history, a backward recurrent state encoding the source future, and a third LSTM layer to combine them. All layers have 350 nodes. The neural networks are implemented using an extension of the RWTHLM toolkit (Sundermeyer et al., 2014b). The parameter weights are optimized with MERT (Och, 2003) towards the BLEU metric.

3.6 RWTH Neural System

The second system provided by the RWTH is an attention-based recurrent neural network similar to (Bahdanau et al., 2015). The implementation is based on Blocks (van Merriënboer et al., 2015) and Theano (Bergstra et al., 2010; Bastien et al., 2012).

The network uses the 30K most frequent words on the source and target side as input vocabulary. The decoder and encoder word embeddings are of size 620. The encoder uses a bidirectional layer with 1024 GRUs (Cho et al., 2014) to encode the source side, while the decoder uses 1024 GRU layer.

The network is trained for up to 300K updates with a minibatch size of 80 using Adadelta (Zeiler, 2012). The network is evaluated every 10000 updates on BLEU and the best network on the newsdev2016/1 dev set is selected as the final network.

The monolingual News Crawl 2015 corpus is translated into English with a simple phrase-based translation system to create additional parallel training data. The new data is weighted by using the News Crawl 2015 corpus (2.3M sentences) once, the Europarl corpus (0.4M sentences) twice and the SETimes2 corpus (0.2M sentences) three times. The final system is an ensemble of 4 networks, all with the same configuration and training settings.

3.7 Tilde

The Tilde system is a phrase-based machine translation system built on LetsMT infrastructure (Vasijevs et al., 2012) that features language-specific data filtering and cleaning modules. Tilde's system was trained on all available parallel data. Two language models are trained using KenLM (Heafield, 2011): 1) a 5-gram model using the Europarl and SETimes2 corpora, and 2) a 3-gram model using the Common Crawl corpus. We also apply a custom tokenization tool that takes into account specifics of the Romanian language and handles non-translatable entities (e.g., file paths, URLs, e-mail addresses, etc.). During translation a rule-based localisation feature is applied.

3.8 Edinburgh/LMU Hierarchical System

The UEDIN-LMU HPBT system is a hierarchical phrase-based machine translation system (Chiang, 2005) built jointly by the University of Edinburgh and LMU Munich. The system is based on the open source Moses implementation of the hierarchical phrase-based paradigm (Hoang et al., 2009). In addition to a set of standard features in a log-linear combination, a number of non-standard enhancements are employed to achieve improved translation quality.

Specifically, we integrate individual language models trained over the separate corpora (News Crawl 2015, Europarl, SETimes2) directly into the log-linear combination of the system and let MIRA (Cherry and Foster, 2012) optimize their weights along with all other features in tuning, rather than relying on a single linearly interpolated language model. We add another background language model estimated over a concatenation of all Romanian corpora including Common Crawl. All language models are unpruned.

For hierarchical rule extraction, we impose less strict extraction constraints than the Moses defaults. We extract more hierarchical rules by allowing for a maximum of ten symbols on the source side, a maximum span of twenty words, and no lower limit to the amount of words covered by right-hand side non-terminals at extraction time. We discard rules with non-terminals on their right-hand side if they are singletons in the training data.

In order to promote better reordering decisions, we implemented a feature in Moses that resembles the phrase orientation model for hierarchical machine translation as described by Huck et al. (2013) and extend our system with it. The model scores orientation classes (*monotone*, *swap*, *discontinuous*) for each rule application in decoding.

We finally follow the approach outlined by Huck et al. (2011) for lightly-supervised training of hierarchical systems. We automatically translate parts (1.2M sentences) of the monolingual Romanian News Crawl 2015 corpus to English with a Romanian→English phrase-based statistical machine translation system (Williams et al., 2016). The foreground phrase table extracted from the human-generated parallel data is filled up with entries from a background phrase table extracted from the automatically produced News Crawl 2015 parallel data.

Huck et al. (2016) give a more in-depth description of the Edinburgh/LMU hierarchical machine translation system, along with detailed experimental results.

3.9 Edinburgh Neural System

Edinburgh's neural machine translation system is an attentional encoder-decoder (Bahdanau et al., 2015), which we train with nematus.[3] We use byte-pair-encoding (BPE) to achieve open-vocabulary translation with a fixed vocabulary of subword symbols (Sennrich et al., 2016c). We produce additional parallel training data by automatically translating the monolingual Romanian News Crawl 2015 corpus into English (Sennrich et al., 2016b), which we combine with the original parallel data in a 1-to-1 ratio. We use minibatches of size 80, a maximum sentence length of 50, word embeddings of size 500, and hidden layers of size 1024. We apply dropout to all layers (Gal, 2015), with dropout probability 0.2, and also drop out full words with probability 0.1. We clip the gradient norm to 1.0 (Pascanu et al., 2013). We train the models with Adadelta (Zeiler, 2012), reshuffling the training corpus between epochs. We validate the model every 10 000 minibatches via BLEU on a validation set, and perform early stopping on BLEU. Decoding is performed with beam search with a beam size of 12.

A more detailed description of the system, and more experimental results, can be found in (Sennrich et al., 2016a).

3.10 Edinburgh Phrase-based System

Edinburgh's phrase-based system is built using the Moses toolkit, with fast_align (Dyer et al., 2013) for word alignment, and KenLM (Heafield et al., 2013) for language model training. In our Moses setup, we use hierarchical lexicalized reordering (Galley and Manning, 2008), operation sequence model (Durrani et al., 2013), domain indicator features, and binned phrase count features. We use all available parallel data for the translation model, and all available Romanian text for the language model. We use two different 5-gram language models; one built from all the monolingual target text concatenated, without pruning, and one built from only News Crawl 2015, with singleton 3-grams and above pruned out. The weights of all these features and models are tuned with k-best MIRA (Cherry and Foster, 2012) on first the half of newsdev2016. In decoding, we use MBR (Kumar and Byrne, 2004), cube-pruning (Huang and Chiang, 2007) with a pop-limit of 5000, and the Moses "monotone at punctuation" switch (to prevent reordering across punctuation) (Koehn and Haddow, 2009).

3.11 USFD Phrase-based System

USFD's phrase-based system is built using the Moses toolkit, with MGIZA (Gao and Vogel, 2008) for word alignment and KenLM (Heafield et al., 2013) for language model training. We use all available parallel data for the translation model. A single 5-gram language model is built using all the target side of the parallel data and a subpart of the monolingual Romanian corpora selected with Xenc-v2 (Rousseau, 2013). For the latter we use all the parallel data as in-domain data and the first half of newsdev2016 as development set. The feature weights are tuned with MERT (Och, 2003) on the first half of newsdev2016.

The system produces distinct 1000-best lists, for which we extend the feature set with the 17 baseline *black-box* features from sentence-level Quality Estimation (QE) produced with Quest++[4] (Specia et al., 2015). The 1000-best lists are then reranked and the top-best hypothesis extracted using the nbest rescorer available within the Moses toolkit.

3.12 UvA

We use a phrase-based machine translation system (Moses) with a distortion limit of 6 and lexicalized reordering. Before translation, the English source side is preordered using the neural preordering model of (de Gispert et al., 2015). The preordering model is trained for 30 iterations on the full MGIZA-aligned training data. We use two language models, built using KenLM. The first is a 5-gram language model trained on all available data. Words in the Common Crawl dataset that appear fewer than 500 times were replaced by UNK, and all singleton ngrams of order 3 or higher were pruned. We also use a 7-gram class-based language model, trained on the same data. 512 word

[3] https://github.com/rsennrich/nematus

[4] http://www.quest.dcs.shef.ac.uk/quest_files/features_blackbox_baseline_17

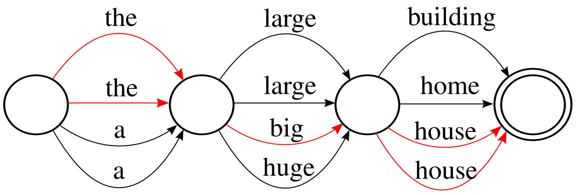

Figure 1: System A: *the large building*; System B: *the large home*; System C: *a big house*; System D: *a huge house*; Reference: *the big house*.

classes were generated using the method of Green et al. (2014).

4 System Combination

System combination produces consensus translations from multiple hypotheses which are obtained from different translation approaches, i.e., the systems described in the previous section. A system combination implementation developed at RWTH Aachen University (Freitag et al., 2014a) is used to combine the outputs of the different engines. The consensus translations outperform the individual hypotheses in terms of translation quality.

The first step in system combination is the generation of confusion networks (CN) from I input translation hypotheses. We need pairwise alignments between the input hypotheses, which are obtained from METEOR (Banerjee and Lavie, 2005). The hypotheses are then reordered to match a selected skeleton hypothesis in terms of word ordering. We generate I different CNs, each having one of the input systems as the skeleton hypothesis, and the final lattice is the union of all I generated CNs. In Figure 1 an example of a confusion network with $I = 4$ input translations is depicted. Decoding of a confusion network finds the best path in the network. Each arc is assigned a score of a linear model combination of M different models, which includes word penalty, 3-gram language model trained on the input hypotheses, a binary primary system feature that marks the primary hypothesis, and a binary voting feature for each system. The binary voting feature for a system is 1 if and only if the decoded word is from that system, and 0 otherwise. The different model weights for system combination are trained with MERT (Och, 2003).

5 Experimental Evaluation

Since only one development set was provided we split the given development set into two parts: newsdev2016/1 and newsdev2016/2. The first part was used as development set while the second part was our internal test set. Additionally we extracted 2000 sentences from the Europarl and SETimes2 data to create two additional development and test sets. Most single systems are optimized for newsdev2016/1 and/or the SETimes2 test set. The system combination was optimized on the newsdev2016/1 set.

The single system scores in Table 1 show clearly that the UEDIN NMT system is the strongest single system by a large margin. The other standalone attention-based neural network contribution, RWTH NMT, follows, with only a small margin before the phrase-based contributions. The combination of all systems improved the strongest system by another 1.9 BLEU points on our internal test set, newsdev2016/2, and by 1 BLEU point on the official test set, newstest2016.

Removing the strongest system from our system combination shows a large degradation of the results. The combination is still slightly stronger then the UEDIN NMT system on newsdev2016/2, but lags behind on newstest2016. Removing the by itself weakest system shows a slight degradation on newsdev2016/2 and newstest2016, hinting that it still provides valuable information.

Table 2 shows a comparison between all systems by scoring the translation output against each other in TER and BLEU. We see that the neural networks outputs differ the most from all the other systems.

6 Morphology Prediction Precision

In order to assess how well the different system outputs predict the right morphology, we compute a precision rate for each Romanian morphological attribute that occurs with nouns, pronouns, adjectives, determiners, and verbs (Table 3). For this purpose, we use the METEOR toolkit (Banerjee and Lavie, 2005) to obtain word alignments between each system translation and the reference translation for newstest2016. The reference and hypotheses are tagged with TTL (Tufiş et al., 2008).[5] Each word in the reference that is assigned a POS tag of interest (noun, pronoun, adjective, determiner, or verb) is then compared to the word it is aligned to in the system output. When, for

[5] The hypotheses were tagged despite the risks that go along with tagging automatically generated sentences. A dictionary would have been a solution, but unfortunately we had no such resource for Romanian.

Individual Systems	newsdev2016/1 BLEU	newsdev2016/1 TER	newsdev2016/2 BLEU	newsdev2016/2 TER	newstest2016 BLEU	newstest2016 TER
KIT	25.2	57.5	29.9	51.8	26.3	55.9
LIMSI	23.3	59.5	27.2	55.0	23.9	59.2
LMU-CUNI	23.4	60.4	28.4	53.5	24.7	58.1
LMU	23.3	60.5	28.6	53.8	24.5	58.5
RWTH HPBT	25.4	58.7	29.3	53.3	25.9	57.6
RWTH NMT	25.1	57.4	30.6	49.6	26.5	55.4
Tilde	21.3	62.7	25.8	56.3	23.2	60.2
UEDIN-LMU HPBT	24.8	58.7	30.1	52.3	25.4	57.7
UEDIN PBT	24.7	59.3	29.1	53.2	25.2	58.1
UEDIN NMT	26.8	56.1	31.4	50.3	27.9	54.5
USFD	22.9	60.4	27.8	54.0	24.4	58.5
UvA	22.1	61.0	27.7	54.2	24.1	58.7
System Combination	28.7	55.5	33.3	49.0	28.9	54.2
- without UEDIN NMT	27.4	56.6	31.6	50.9	27.5	55.4
- without Tilde	28.8	55.5	33.0	49.5	28.7	54.5

Table 1: Results of the individual systems for the English→Romanian task. BLEU [%] and TER [%] scores are case-sensitive.

	KIT	LIMSI	LMU-CUNI	LMU	RWTH HPBT	RWTH NMT	Tilde	UEDIN-LMU HPBT	UEDIN PBT	UEDIN NMT	USFD	UvA	Average
KIT	-	55.0	55.9	51.7	56.2	48.2	50.3	54.6	55.1	42.8	56.6	54.1	52.8
LIMSI	29.3	-	54.3	52.1	51.8	43.0	49.8	55.3	56.2	38.2	57.3	52.1	51.4
LMU-CUNI	28.5	30.8	-	52.4	53.3	43.8	55.4	56.0	56.6	39.3	58.6	56.6	52.9
LMU	31.2	32.0	31.7	-	53.6	43.1	49.1	59.4	58.6	37.8	56.1	55.8	51.8
RWTH HPBT	28.5	32.4	31.2	30.8	-	47.5	50.1	54.9	55.6	41.8	53.9	55.3	52.2
RWTH NMT	33.7	37.9	37.3	37.5	34.8	-	40.8	44.3	45.3	46.0	43.8	43.6	44.5
Tilde	32.2	33.7	29.6	33.8	33.4	39.6	-	53.4	58.5	36.5	55.5	52.0	50.1
UEDIN-LMU HPBT	29.5	29.9	29.4	27.3	29.8	36.9	30.9	-	62.8	38.9	59.6	56.2	54.1
UEDIN PBT	28.4	28.9	28.5	27.0	29.3	35.4	27.0	24.2	-	39.4	60.2	58.6	55.2
UEDIN NMT	38.6	42.6	42.0	43.0	40.1	35.5	44.0	42.1	41.1	-	38.2	38.2	39.7
USFD	27.6	28.8	27.4	28.8	30.4	37.0	29.1	26.5	25.7	42.6	-	58.8	54.4
UvA	29.9	32.0	28.6	29.2	29.6	37.5	31.5	29.0	26.5	43.2	26.9	-	52.9
Average	30.7	32.6	31.4	32.0	31.8	36.6	33.2	30.5	29.3	41.3	30.0	31.3	-

Table 2: Comparison of system outputs against each other, generated by computing BLEU and TER on the system translations for newstest2016. One system in a pair is used as the reference, the other as candidate translation; we report the average over both directions. The upper-right half lists BLEU [%] scores, the lower-left half TER [%] scores.

Attribute	KIT	LIMSI	LMU-CUNI	LMU	RWTH HPBT	RWTH NMT	Tilde	UEDIN-LMU HPBT	UEDIN PBT	UEDIN NMT	USFD	UvA	Combination
Case	46.7%	46.0%	46.3%	45.7%	47.7%	48.0%	44.4%	46.3%	47.4%	49.8%	45.4%	45.4%	**50.8%**
Definite	50.5%	49.1%	50.0%	49.2%	50.5%	50.1%	47.2%	50.0%	50.5%	51.0%	49.2%	48.9%	**53.3%**
Gender	51.9%	51.0%	51.9%	51.3%	52.6%	52.1%	49.6%	51.9%	52.7%	53.0%	51.2%	50.9%	**54.9%**
Number	53.2%	51.7%	52.6%	52.3%	53.6%	53.7%	50.6%	52.9%	53.6%	54.9%	52.1%	51.8%	**56.3%**
Person	52.8%	51.3%	52.0%	52.0%	53.5%	55.0%	50.6%	52.6%	53.4%	**57.2%**	52.4%	51.6%	57.1%
Tense	45.8%	44.1%	44.7%	44.8%	45.7%	45.5%	42.3%	45.2%	45.1%	46.6%	44.9%	44.8%	**48.0%**
Verb form	45.9%	44.4%	45.5%	44.9%	46.6%	47.0%	43.9%	46.1%	46.5%	47.2%	45.5%	43.3%	**48.7%**
Reference words with alignment	57.7%	56.7%	57.3%	57.3%	58.3%	57.6%	55.7%	58.0%	58.5%	58.3%	57.3%	56.8%	**60.4%**

Table 3: Precision of each system on morphological attribute prediction computed over the reference translation using METEOR alignments. The last row shows the ratio of reference words for which METEOR managed to find an alignment in the hypothesis.

a given morphological attribute, the output and the reference have the same value (e.g. *Number=Singular*), we consider the prediction correct. The prediction is considered wrong in every other case.

The last row in Table 3 shows the ratio of reference words for which METEOR found an alignment in the hypothesis. We observe a high correlation between this ratio and the quality of the morphological predictions, showing that the accuracy is highly dependent on the alignments. We nevertheless observe that the predictions made by UEDIN NMT are strictly all better than UEDIN PBT, although the latter has slightly more alignments to the reference. The system combination makes the most accurate predictions for almost every attribute. The difference in precision with the best single system (UEDIN NMT) can be significant (2.3% for definite and 1.4% for tense) showing that the combination managed to effectively identify the strong points of each translation system.

7 Conclusion

Our combined effort shows that even with an extremely strong single best system, we still manage to improve the final result by one BLEU point by combining it with the other systems of all participating research groups.

The joint submission for English→Romanian is the best submission measured in terms of BLEU, as presented on the WMT submission page.[6]

[6] http://matrix.statmt.org/

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements № 645452 (*QT21*) and 644402 (*HimL*).

References

Alexandre Allauzen, Lauriane Aufrant, Franck Burlot, Elena Knyazeva, Thomas Lavergne, and François Yvon. 2016. LIMSI@WMT'16 : Machine translation of news. In *Proc. of the ACL 2016 First Conf. on Machine Translation (WMT16)*, Berlin, Germany, August.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72, Ann Arbor, MI, USA, June.

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. 2012. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June. Oral Presentation.

Fabienne Braune, Nina Seemann, and Alexander Fraser. 2015. Rule Selection with Soft Syntactic Features for String-to-Tree Statistical Machine Translation. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*.

Fabienne Braune, Alexander Fraser, Hal Daumé III, and Aleš Tamchyna. 2016. A Framework for Discriminative Rule Selection in Hierarchical Moses. In *Proc. of the ACL 2016 First Conf. on Machine Translation (WMT16)*, Berlin, Germany, August.

Francesco Casacuberta and Enrique Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(3):205–225.

Boxing Chen, Roland Kuhn, George Foster, and Howard Johnson. 2011. Unpacking and Transforming Feature Functions: New Ways to Smooth Phrase Tables. In *MT Summit XIII*, pages 269–275, Xiamen, China, September.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proc. of the Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 427–436, Montréal, Canada, June.

David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 263–270, Ann Arbor, Michigan, June.

David Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Joseph M. Crego and José B. Mariño. 2006. Improving statistical MT by coupling reordering and decoding. *Machine translation*, 20(3):199–215, Jul.

Josep Maria Crego, Franois Yvon, and José B. Mariño. 2011. N-code: an open-source Bilingual N-gram SMT Toolkit. *Prague Bulletin of Mathematical Linguistics*, 96:49–58.

Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2015. Fast and accurate preordering for SMT using neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1012–1017, Denver, Colorado, May–June.

Nadir Durrani, Alexander Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. 2013. Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT? In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 399–405, Sofia, Bulgaria, August.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of NAACL-HLT*, pages 644–648, Atlanta, Georgia, June.

Markus Freitag, Stephan Peitz, Joern Wuebker, Hermann Ney, Nadir Durrani, Matthias Huck, Philipp Koehn, Thanh-Le Ha, Jan Niehues, Mohammed Mediani, Teresa Herrmann, Alex Waibel, Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 128–135, Heidelberg, Germany, December.

Markus Freitag, Matthias Huck, and Hermann Ney. 2014a. Jane: Open Source Machine Translation System Combination. In *Proc. of the Conf. of the European Chapter of the Assoc. for Computational Linguistics (EACL)*, pages 29–32, Gothenberg, Sweden, April.

Markus Freitag, Stephan Peitz, Joern Wuebker, Hermann Ney, Matthias Huck, Rico Sennrich, Nadir Durrani, Maria Nadejde, Philip Williams, Philipp Koehn, Teresa Herrmann, Eunah Cho, and Alex Waibel. 2014b. EU-BRIDGE MT: Combined Machine Translation. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 105–113, Baltimore, MD, USA, June.

Markus Freitag, Joern Wuebker, Stephan Peitz, Hermann Ney, Matthias Huck, Alexandra Birch, Nadir Durrani, Philipp Koehn, Mohammed Mediani, Isabel Slawik, Jan Niehues, Eunah Cho, Alex Waibel, Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2014c. Combined Spoken Language Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 57–64, Lake Tahoe, CA, USA, December.

Yarin Gal. 2015. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. *ArXiv e-prints*.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57. Association for Computational Linguistics.

Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with LSTM. *Neural computation*, 12(10):2451–2471.

Felix A. Gers, Nicol N. Schraudolph, and Jürgen Schmidhuber. 2003. Learning precise timing with lstm recurrent networks. *The Journal of Machine Learning Research*, 3:115–143.

Joshua Goodman. 2001. Classes for fast maximum entropy training. *CoRR*, cs.CL/0108006.

Spence Green, Daniel Cer, and Christopher D. Manning. 2014. An empirical comparison of features and tuning for phrase-based machine translation. In *In Procedings of the Ninth Workshop on Statistical Machine Translation*.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. pages 690–696, Sofia, Bulgaria, August.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

Teresa Herrmann, Jan Niehues, and Alex Waibel. 2015. Source Discriminative Word Lexicon for Translation Disambiguation. In *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT15)*, Danang, Vietnam.

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. A Unified Framework for Phrase-Based, Hierarchical, and Syntax-Based Statistical Machine Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 152–159, Tokyo, Japan, December.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Liang Huang and David Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 144–151, Prague, Czech Republic, June.

Matthias Huck, David Vilar, Daniel Stein, and Hermann Ney. 2011. Lightly-Supervised Training for Hierarchical Phrase-Based Machine Translation. In *Proc. of the EMNLP 2011 Workshop on Unsupervised Learning in NLP*, pages 91–96, Edinburgh, Scotland, UK, July.

Matthias Huck, Joern Wuebker, Felix Rietig, and Hermann Ney. 2013. A Phrase Orientation Model for Hierarchical Machine Translation. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 452–463, Sofia, Bulgaria, August.

Matthias Huck, Alexander Fraser, and Barry Haddow. 2016. The Edinburgh/LMU Hierarchical Machine Translation System for WMT 2016. In *Proc. of the ACL 2016 First Conf. on Machine Translation (WMT16)*, Berlin, Germany, August.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving Translation Quality by Discarding Most of the Phrasetable. In *Proc. of EMNLP-CoNLL 2007*.

Philipp Koehn and Barry Haddow. 2009. Edinburgh's Submission to all Tracks of the WMT 2009 Shared Task with Reordering and Speed Improvements to Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 160–164, Athens, Greece.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantine, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. pages 177–180, Prague, Czech Republic, June.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-Risk Decoding for Statistical Machine Translation. In *HLT 2004 - Human Language Technology Conference*, Boston, MA, May.

José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, and Marta R. Costa-jussà. 2006. N-gram-based machine translation. *Comput. Linguist.*, 32(4):527–549, December.

R.C. Moore and W. Lewis. 2010. Intelligent Selection of Language Model Training Data. In *ACL (Short Papers)*, pages 220–224, Uppsala, Sweden, July.

Frederic Morin and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 246–252. Society for Artificial Intelligence and Statistics.

J. Niehues and A. Waibel. 2012. Detailed Analysis of Different Strategies for Phrase Table Adaptation in SMT. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas*, San Diego, CA, USA.

Jan Niehues and Alex Waibel. 2013. An MT Error-Driven Discriminative Word Lexicon using Sentence Structure Features. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, Sofia, Bulgaria.

Jan Niehues, Quoc Khanh Do, Alexandre Allauzen, and Alex Waibel. 2015. Listnet-based MT Rescoring. *EMNLP 2015*, page 248.

Franz Josef Och. 1999. An Efficient Method for Determining Bilingual Word Classes. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*, pages 1310–1318, , Atlanta, GA, USA.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proc. of the 21st Int. Conf. on Computational Linguistics and the 44th Annual Meeting of the Assoc. for Computational Linguistics*, pages 433–440.

Kay Rottmann and Stephan Vogel. 2007. Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, Skövde, Sweden.

Anthony Rousseau. 2013. Xenc: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, (100):73–82.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation (WMT16)*, Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin, Germany.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China, July. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Speech and Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO, September.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM Neural Networks for Language Modeling. In *Interspeech*, Portland, OR, USA, September.

Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker, and Hermann Ney. 2014a. Translation Modeling with Bidirectional Recurrent Neural Networks. In *Conference on Empirical Methods in Natural Language Processing*, pages 14–25, Doha, Qatar, October.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2014b. rwthlm - The RWTH Aachen University Neural Network Language Modeling Toolkit . In *Interspeech*, pages 2093–2097, Singapore, September.

Aleš Tamchyna, Fabienne Braune, Alexander M. Fraser, Marine Carpuat, Hal Daumé III, and Chris Quirk. 2014. Integrating a Discriminative Classifier into Phrase-based and Hierarchical Decoding. *The Prague Bulletin of Mathematical Linguistics (PBML)*, 101:29–42.

Aleš Tamchyna, Alexander Fraser, Ondřej Bojar, and Marcin Junczsys-Dowmunt. 2016. Target-Side Context for Discriminative Models in Statistical Machine Translation. In *Proc. of ACL*, Berlin, Germany, August. Association for Computational Linguistics.

Dan Tufiş, Radu Ion, Ru Ceauşu, and Dan Ştefănescu. 2008. RACAI's Linguistic Web Services. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Bart van Merriënboer, Dzmitry Bahdanau, Vincent Dumoulin, Dmitriy Serdyuk, David Warde-Farley, Jan Chorowski, and Yoshua Bengio. 2015. Blocks and Fuel: Frameworks for deep learning. *CoRR*, abs/1506.00619.

Andrejs Vasijevs, Raivis Skadiš, and Jörg Tiedemann. 2012. LetsMT!: A Cloud-Based Platform for Do-It-Yourself Machine Translation. In Min Zhang, editor, *Proceedings of the ACL 2012 System Demonstrations*, number July, pages 43–48, Jeju Island, Korea. Association for Computational Linguistics.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 2010. Jane: Open source hierarchical translation, extended with reordering and lexicon models. In *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 262–270, Uppsala, Sweden, July.

Philip Williams, Rico Sennrich, Maria Nădejde, Matthias Huck, Barry Haddow, and Ondřej Bojar. 2016. Edinburgh's Statistical Machine Translation

Systems for WMT16. In *Proc. of the ACL 2016 First Conf. on Machine Translation (WMT16)*, Berlin, Germany, August.

Joern Wuebker, Stephan Peitz, Felix Rietig, and Hermann Ney. 2013. Improving statistical machine translation with word class models. In *Conference on Empirical Methods in Natural Language Processing*, pages 1377–1381, Seattle, WA, USA, October.

Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *CoRR*, abs/1212.5701.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-Based Statistical Machine Translation. In *25th German Conf. on Artificial Intelligence (KI2002)*, pages 18–32, Aachen, Germany, September. Springer Verlag.